

Causal Rate Distortion Function on Abstract Alphabets and Optimal Reconstruction Kernel

Charalambos D. Charalambous, Photios A. Stavrou and Christos K. Kourtellaris

Abstract—A Causal rate distortion function with a general fidelity criterion is formulated on abstract alphabets and the optimal reconstruction kernel is derived, which consists of a product of causal kernels. In the process, general abstract spaces are introduced to show existence of the minimizing kernel using weak*-convergence. Certain properties of the causal rate distortion function are presented.

I. INTRODUCTION

This paper is concerned with lossy data compression subject to distortion or fidelity criterion and causal decoding on abstract alphabets. Its information theoretic interpretation is the causal rate distortion function formulated via the directed information between the source sequence $X^n \triangleq \{X_0, X_1, \dots, X_n\}$ and its reproduction sequence $Y^n \triangleq \{Y_0, Y_1, \dots, Y_n\}$ defined by

$$I(X^n \rightarrow Y^n) \triangleq \sum_{i=0}^n I(X^i; Y_i | Y^{i-1}) \quad (1)$$

The average distortion constraint is

$$E\{d_{0,n}(X^n, Y^n)\} \leq D, \quad d_{0,n}(x^n, y^n) \triangleq \sum_{i=0}^n \rho_{0,i}(x^i, y^i) \quad (2)$$

where $D \geq 0$, $d_{0,n}(\cdot, \cdot)$ a non-negative distortion function. Define the causal product of conditional distributions by

$$\vec{P}_{Y^n|X^n}(dy^n|x^n) \triangleq \otimes_{i=0}^n P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i) \quad (3)$$

where $P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i)$ denotes the conditional distribution of Y_i given (Y^{i-1}, X^i) , $i = 0, 1, \dots, n$. Since causal codes as defined in [4] satisfy $P_{X_i|X^{i-1}, Y^{i-1}}(dx_i|x^{i-1}, y^{i-1}) = P_{X_i|X^{i-1}}(dx_i|x^{i-1})$. $P - a.s$ (see also Lemma 2.4), in the analysis it is convenient to express $I(X^n \rightarrow Y^n)$ as a functional of $\vec{P}_{Y^n|X^n}(dy^n|x^n)$ as follows.

$$I(X^n \rightarrow Y^n) = \int \log\left(\frac{\vec{P}_{Y^n|X^n}(dy^n|x^n)}{P_{Y^n}(dy^n)}\right) \times \vec{P}_{Y^n|X^n}(dy^n|x^n) P_{X^n}(dx^n) \quad (4)$$

$$= \mathbb{I}(P_{X^n}, \vec{P}_{Y^n|X^n}) \quad (5)$$

where $\mathbb{I}(P_{X^n}, \vec{P}_{Y^n|X^n})$ indicates the functional dependence of $I(X^n \rightarrow Y^n)$ on $\{P_{X^n}, \vec{P}_{Y^n|X^n}\}$.

C. D. Charalambous (chadcha@ucy.ac.cy).

P. A. Stavrou (stavrou.fotios@ucy.ac.cy).

C. K. Kourtellaris (kourtellaris.christos@ucy.ac.cy).

The authors are with the Department of Electrical and Computer Engineering, University of Cyprus, Nicosia, CYPRUS

The causal information rate distortion function investigated is

$$\inf_{\vec{P}_{Y^n|X^n}(dy^n|x^n): E\{d_{0,n}(X^n, Y^n)\} \leq D} I(X^n \rightarrow Y^n) \quad (6)$$

Under appropriate assumptions on $d_{0,n}(\cdot, \cdot)$ it is shown that the optimal causal product (reproduction channel) $\vec{P}_{Y^n|X^n}^*$ which achieves the infimum in (6) is given by

$$\vec{P}_{Y^n|X^n}^*(dy^n|x^n) = \otimes_{i=0}^n \frac{e^{s\rho_i(x^i, y^i)} P_{Y_i|Y^{i-1}}^*(dy_i|y^{i-1})}{\int_{\mathcal{Y}_i} e^{s\rho_i(x^i, y^i)} P_{Y_i|Y^{i-1}}^*(dy_i|y^{i-1})} \quad (7)$$

where $s \leq 0$ is the Lagrange multiplier associated with the fidelity constraint. The operational meaning of (6) is shown in [5] via coding theorems (called sequential code), hence this aspect will not be discussed. Rather, the main emphasis of the paper is the mathematical formulation, the prove of existence of solution to (6), the derivation of (7), the derivation of a closed form expression for the causal rate distortion function, and some of its properties.

The Shannon source code consists of an encoder-decoder pair. The encoder observes a source sequence $X^\infty \triangleq \{X_0, X_1, \dots\}$ and generates a compressed representation $\{Z_0, Z_1, \dots\}$. The decoder upon observing the representation sequence $\{Z_0, Z_1, \dots\}$ generates a reproduction sequence $Y_i = f_i(X^\infty)$ of X_i , for every time step i . The dependence of the reproduction sequence on the future source symbols, in addition to its past and present symbols makes such a decoder non-causal. In Neuhoff and Gilbert [4], a source code is defined as causal if the reproduction sequence is such that $f_i(X^\infty) = f_i(\tilde{X}^\infty)$ whenever $X^i = \tilde{X}^i$, $\forall i = 0, 1, \dots$. The definition of a causal code necessitates that any information theoretic causal rate distortion function should lead to an optimal reconstruction conditional distribution which is causally dependent on the source symbols, and (7) has this property.

The classical rate distortion function is defined via the mutual information between X^n and Y^n , namely, $I(X^n; Y^n)$ with average distortion (2), and the code is assumed non-causal, leading to the well known optimal reconstruction [1], [3]

$$P_{Y^n|X^n}^*(dy^n|x^n) = \frac{e^{s \sum_{i=0}^n \rho_{0,i}(x^i, y^i)} P_{Y^n}^*(dy^n)}{\int_{\mathcal{Y}_{0,n}} e^{s \sum_{i=0}^n \rho_{0,i}(x^i, y^i)} P_{Y^n}^*(dy^n)} \quad (8)$$

Since by chain rule $P_{Y^n|X^n}(dy^n|x^n) = \otimes_{i=0}^n P_{Y_i|Y^{i-1}, X^n=x^n}(dy_i|y^{i-1} = y^{i-1}, X^n = x^n)$, the classical rate distortion theory gives a reconstruction $Y_i = y_i$ which depends on future values of the source symbols,

($X_{i+1} = x_{i+1}, \dots, X_n = x_n$) in addition to its past reconstructions $Y^{i-1} = y^{i-1}$, and past and present source symbols $X^i = x^i$. The point to be made here is that, in general, aside from some special examples, such as the i.i.d source and single letter distortion $d_{0,n} = \sum_{i=0}^n \rho_i(x_i, y_i)$ [2] the reconstruction conditional distribution and hence the decoder of the classical rate distortion function is non-causal. On the other hand, a code is causal if the reconstruction distribution is causal.

II. PROBLEM FORMULATION

In this section, we introduce the set up of the problem on discrete time sets $\mathbb{N}^n \triangleq \{0, 1, \dots, n\}$, $n \in \mathbb{N} \triangleq \{0, 1, 2, \dots\}$. Assume all processes are defined on a complete probability space $(\Omega, \mathcal{F}(\Omega), \mathbb{P})$ with filtration $\{\mathcal{F}_t\}_{t \geq 0}$. The source and reconstruction alphabets are sequences of Polish spaces [11] $\{\mathcal{X}_t : t \in \mathbb{N}\}$ and $\{\mathcal{Y}_t : t \in \mathbb{N}\}$, respectively, (e.g., $\mathcal{Y}_t, \mathcal{X}_t$ are complete separable metric spaces), associated with their corresponding measurable spaces $(\mathcal{X}_t, \mathcal{B}(\mathcal{X}_t))$ and $(\mathcal{Y}_t, \mathcal{B}(\mathcal{Y}_t))$ (e.g., $\mathcal{B}(\mathcal{X}_t)$ is a Borel σ -algebra of subsets of the set \mathcal{X}_t generated by closed sets), $t \in \mathbb{N}$. Sequences of alphabets are identified with the product spaces $(\mathcal{X}_{0,n}, \mathcal{B}(\mathcal{X}_{0,n})) \triangleq \times_{k=0}^n (\mathcal{X}_k, \mathcal{B}(\mathcal{X}_k))$, and $(\mathcal{Y}_{0,n}, \mathcal{B}(\mathcal{Y}_{0,n})) \triangleq \times_{k=0}^n (\mathcal{Y}_k, \mathcal{B}(\mathcal{Y}_k))$. The source and reconstruction are processes denoted by $X^n \triangleq \{X_t : t \in \mathbb{N}^n\}$, $X : \mathbb{N}^n \times \Omega \mapsto \mathcal{X}_t$, and by $Y^n \triangleq \{Y_t : t \in \mathbb{N}^n\}$, $Y : \mathbb{N}^n \times \Omega \mapsto \mathcal{Y}_t$, respectively. Probability measures on any measurable space $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ are denoted by $\mathcal{M}_1(\mathcal{Z})$. It is assumed that the σ -algebras $\sigma\{X^{-1}\} = \sigma\{Y^{-1}\} = \{\emptyset, \Omega\}$.

Definition 2.1: Let $(\mathcal{X}, \mathcal{B}(\mathcal{X})), (\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ be measurable spaces in which \mathcal{Y} is a Polish Space.

A stochastic Kernel on \mathcal{Y} given \mathcal{X} is a mapping $q : \mathcal{B}(\mathcal{Y}) \times \mathcal{X} \rightarrow [0, 1]$ satisfying the following two properties:

- 1) For every $x \in \mathcal{X}$, the set function $q(\cdot; x)$ is a probability measure (possibly finitely additive) on $\mathcal{B}(\mathcal{Y})$.
- 2) For every $F \in \mathcal{B}(\mathcal{Y})$, the function $q(F; \cdot)$ is $\mathcal{B}(\mathcal{X})$ -measurable.

The set of all such stochastic Kernels is denoted by $\mathcal{Q}(\mathcal{Y}; \mathcal{X})$.

An important notion is conditional independence. The Random Variable (R.V.) Z is called conditional independent of R.V. X given the R.V. Y if and only if $X \leftrightarrow Y \leftrightarrow Z$ forms a Markov chain in both directions.

Stochastic kernels can be used to define non-causal and causal product reconstruction kernels and associated rate distortion functions.

Definition 2.2: Given measurable spaces $(\mathcal{X}_{0,n}, \mathcal{B}(\mathcal{X}_{0,n}))$, $(\mathcal{Y}_{0,n}, \mathcal{B}(\mathcal{Y}_{0,n}))$, and their product spaces, data compression channels are defined as follows.

- 1) A *Non-Causal Data Compression Channel* is a stochastic kernel $q_{0,n}(dy^n; x^n) \in \mathcal{Q}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$, $n \in \mathbb{N}$.
- 2) A *Causal Product Data Compression Channel* is a product of a sequence of causal stochastic kernels defined by

$$\vec{q}_{0,n}(dy^n; x^n) = \otimes_{i=0}^n q_i(dy_i; y^{i-1}, x^i)$$

where $q_i \in \mathcal{Q}(\mathcal{Y}_i; \mathcal{Y}_{0,i-1} \times \mathcal{X}_{0,i})$, $i = 0, \dots, n$, $n \in \mathbb{N}$.

Note that classical rate distortion theory is concerned with finding the optimal $P_{Y^n|X^n}(dy^n|X^n = x^n)$, which is generally non-causal, while in this paper the interest is to find the optimal causal product kernel.

A. Causal and Classical Rate Distortion Functions

In this section the classical rate distortion function which has a non-causal structure is reviewed, and then the causal rate distortion function is defined.

Given a source probability measure $\mu_{0,n} \in \mathcal{M}_1(\mathcal{X}_{0,n})$ (possibly finite additive) and a reconstruction Kernel $q_{0,n} \in \mathcal{Q}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$, one can define three probability measures as follows.

(P1): The joint measure $P_{0,n} \in \mathcal{M}_1(\mathcal{Y}_{0,n} \times \mathcal{X}_{0,n})$:

$$\begin{aligned} P_{0,n}(G_{0,n}) &\triangleq (\mu_{0,n} \otimes q_{0,n})(G_{0,n}), \quad G_{0,n} \in \mathcal{B}(\mathcal{X}_{0,n}) \times \mathcal{B}(\mathcal{Y}_{0,n}) \\ &= \int_{\mathcal{X}_{0,n}} q_{0,n}(G_{0,n}, x^n; x^n) \mu_{0,n}(dx^n) \end{aligned}$$

where $G_{0,n}, x^n$ is the x^n -section of $G_{0,n}$ at point x^n defined by $G_{0,n}, x^n \triangleq \{y^n \in \mathcal{Y}_{0,n} : (x^n, y^n) \in G_{0,n}\}$ and \otimes denotes the convolution.

(P2): The marginal measure $\nu_{0,n} \in \mathcal{M}_1(\mathcal{Y}_{0,n})$:

$$\begin{aligned} \nu_{0,n}(F_{0,n}) &\triangleq P_{0,n}(\mathcal{X}_{0,n} \times F_{0,n}), \quad F_{0,n} \in \mathcal{B}(\mathcal{Y}_{0,n}) \\ &= \int_{\mathcal{X}_{0,n}} q_{0,n}((\mathcal{X}_{0,n} \times F_{0,n})_{x^n}; x^n) \mu_{0,n}(dx^n) \\ &= \int_{\mathcal{X}_{0,n}} q_{0,n}(F_{0,n}; x^n) \mu_{0,n}(dx^n) \end{aligned}$$

(P3): The product measure $\pi_{0,n} : \mathcal{B}(\mathcal{X}_{0,n}) \times \mathcal{B}(\mathcal{Y}_{0,n}) \mapsto [0, 1]$ of $\mu_{0,n} \in \mathcal{M}_1(\mathcal{X}_{0,n})$ and $\nu_{0,n} \in \mathcal{M}_1(\mathcal{Y}_{0,n})$:

$$\begin{aligned} \pi_{0,n}(G_{0,n}) &\triangleq (\mu_{0,n} \times \nu_{0,n})(G_{0,n}), \quad G_{0,n} \in \mathcal{B}(\mathcal{X}_{0,n}) \times \mathcal{B}(\mathcal{Y}_{0,n}) \\ &= \int_{\mathcal{X}_{0,n}} \nu_{0,n}(G_{0,n}, x^n) \mu_{0,n}(dx^n) \end{aligned}$$

The precise definition of mutual information between two sequences of Random Variables X^n and Y^n , denoted $I(X^n; Y^n)$ is defined via the Kullback-Leibler distance (or relative entropy) between the joint probability distribution of (X^n, Y^n) and the product of its marginal probability distributions of X^n and Y^n , using the Radon-Nikodym derivative. Hence, by the construction of probability measures (P1)-(P3), and the chain rule of relative entropy [11]:

$$I(X^n; Y^n) \triangleq \mathbb{D}(P_{0,n} || \pi_{0,n}) \quad (9)$$

$$\begin{aligned} &= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{d(\mu_{0,n} \otimes q_{0,n})}{d(\mu_{0,n} \times \nu_{0,n})} \right) d(\mu_{0,n} \otimes q_{0,n}) \\ &= \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{q_{0,n}(dy^n; x^n)}{\nu_{0,n}(dy^n)} \right) \\ &\quad q_{0,n}(dy^n; dx^n) \mu_{0,n}(dx^n) \\ &= \int_{\mathcal{X}_{0,n}} \mathbb{D}(q_{0,n}(\cdot; x^n) || \nu_{0,n}(\cdot)) \mu_{0,n}(dx^n) \\ &\equiv \mathbb{I}(\mu_{0,n}; q_{0,n}) \end{aligned} \quad (10)$$

Note that (10) states that mutual information is expressed as a functional of $\{\mu_{0,n}, q_{0,n}\}$ and it is denoted by $\mathbb{I}(\mu_{0,n}; q_{0,n})$.

Note that necessary and sufficient conditions for existence of a Radon-Nikodym derivative for finitely additive measures can be found in [13]. Moreover, $I(X^n; Y^n)$ is also expressed by the sum of two directed information as follows

$$I(X^n; Y^n) = I(X^n \rightarrow Y^n) + I(X^n \leftarrow Y^n) \quad (11)$$

where

$$I(X^n \rightarrow Y^n) \triangleq \sum_{i=0}^n I(X^i; Y_i | Y^{i-1}) \quad (12)$$

$$I(X^n \leftarrow Y^n) \triangleq \sum_{i=0}^n I(Y^{i-1}; X_i | X^{i-1}) \quad (13)$$

Definition 2.3: (Classical Rate Distortion Function) Let $d_{0,n} : \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n} \rightarrow [0, \infty)$, be an $\mathcal{B}(\mathcal{X}_{0,n}) \times \mathcal{B}(\mathcal{Y}_{0,n})$ -measurable distortion function, and let $Q_{0,n}(D) \subset \mathcal{Q}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$ (assuming is non-empty) denotes the average distortion or fidelity constraint defined by

$$Q_{0,n}(D) \triangleq \left\{ q_{0,n} \in \mathcal{Q}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n}) : \frac{1}{n+1} \int_{\mathcal{X}_{0,n}} \int_{\mathcal{Y}_{0,n}} d_{0,n}(x^n, y^n) q_{0,n}(dy^n; x^n) \mu_{0,n}(dx^n) \leq D \right\}, \quad D \geq 0 \quad (14)$$

The classical rate distortion function associated with the non-causal kernel $q_{0,n} \in \mathcal{Q}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$ is defined by

$$R_{0,n}(D) \triangleq \inf_{q_{0,n} \in Q_{0,n}(D)} \frac{1}{n+1} \mathbb{I}(\mu_{0,n}; q_{0,n}) \quad (15)$$

while its operational meaning can be established via $\limsup_{n \rightarrow \infty} R_{0,n}$.

Existence in (15) is shown assuming $d_{0,n}(x^n; \cdot)$ is bounded continuous on $\mathcal{Y}_{0,n}$ and $\mathcal{Y}_{0,n}$ is compact, using weak-convergence of probability measures in [3], and for more general $d_{0,n}(x^n; \cdot)$ which is only continuous in $\mathcal{Y}_{0,n}$ using weak*-convergence of measures [14] on Polish spaces.

A version of the optimal reconstruction kernel which attains the infimum in (15), [3] is

$$q_{0,n}^*(dy^n; x^n) = \frac{e^{s d_{0,n}(x^n, y^n)} \nu_{0,n}^*(dy^n)}{\int_{\mathcal{Y}_{0,n}} e^{s d_{0,n}(x^n, y^n)} \nu_{0,n}^*(dy^n)}, \quad s \leq 0 \quad (16)$$

where $\nu_{0,n}^* \in \mathcal{M}_1(\mathcal{Y}_{0,n})$ is the marginal of $P_{0,n}^* = \mu_{0,n} \otimes q_{0,n}^* \in \mathcal{M}_1(\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n})$ and $s \leq 0$ is the Lagrange multiplier associated with the fidelity constraint (14). Unfortunately, for general sources and distortion function $d_{0,n}$, the optimal reconstruction $q_{0,n}^*(dy^n; x^n) = \otimes_{i=0}^n q_i^*(dy_i; y^{i-1}, x^n)$ is non-causal and introduces delay in the reconstruction processes. On the other hand, if the solution (16) gives a reconstruction such that $q_{0,n}^*(dy^n; x^n) = \vec{q}_{0,n}^*(dy^n; x^n) = \otimes_{i=0}^n q_i^*(dy_i; y^{i-1}, x^i)$ it will be causal. However, there are only limited examples in which (16) is causal on the source sequence. For single letter distortion function $d_{0,n}(x^n, y^n) = \frac{1}{n+1} \sum_{i=0}^n \rho_i(x_i, y_i)$ and independent sources $\mu_{0,n}(dx^n) = \otimes_{i=0}^n \mu_i(dx^i)$ (e.g., $\{X_i : i \in \mathbb{N}\}$ are independent) the optimal reconstruction $q_{0,n}^*(dy^n; x^n)$ factors into a product of causal kernels $q_{0,n}^*(dy^n; x^n) = \otimes_{i=0}^n q_i(dy_i, x_i)$ [2].

This raises the question whether the classical rate distortion function can be reformulated using the causal product $\vec{q}_{0,n}(dy^n; x^n)$.

The next lemma relates causal product reconstruction kernels, mutual information, directed information, and conditional independence.

Lemma 2.4: The following are equivalent for each $n \in \mathbb{N}$.

- 1) $q_{0,n}(dy^n; x^n) = \vec{q}_{0,n}(dy^n; x^n)$, as defined in Definition 2.2-2)
- 2) For each $i = 0, 1, \dots, n-1$, $Y_i \leftrightarrow (X^i, Y^{i-1}) \leftrightarrow (X_{i+1}, X_{i+2}, \dots, X_n)$, forms a Markov chain
- 3) $I(X^n; Y^n) = I(X^n \rightarrow Y^n)$
- 4) $I(X^n \leftarrow Y^n) = 0$
- 5) For each $i = 0, 1, \dots, n-1$, $Y^i \leftrightarrow X^i \leftrightarrow X_{i+1}$ forms a Markov chain

Proof. Omitted due to space limitation.

According to Lemma 2.4 any source with a satisfying conditional distribution $P_{X_i | X^{i-1}, Y^{i-1}}(dx_i | X^{i-1} = x^{i-1}, Y^{i-1} = y^{i-1}) = P_{X_i | X^{i-1}}(dx_i | X^{i-1} = x^{i-1})$, $P - a.s.$, $\forall i \in \mathbb{N}$ is equivalent to any of the equivalent statements of Lemma 2.4. Therefore, for such a source the mutual information becomes

$$I(X^n; Y^n) = I(X^n \rightarrow Y^n) = \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \left(\frac{\vec{q}_{0,n}(dy^n; x^n)}{\nu_{0,n}(dy^n)} \right) \vec{q}_{0,n}(dy^n; dx^n) \mu_{0,n}(dx^n) \quad (17)$$

$$\equiv \mathbb{I}(\mu_{0,n}; \vec{q}_{0,n}) \quad (18)$$

where (18) states that $I(X^n; Y^n)$ is a functional of $\{\mu_{0,n}, \vec{q}_{0,n}\}$. Hence, causal rate distortion is defined by optimizing $\mathbb{I}(\mu_{0,n}; \vec{q}_{0,n})$ over $\vec{q}_{0,n}$ which satisfies a distortion constraint.

Definition 2.5: (Causal Rate Distortion Function) Suppose $d_{0,n} \triangleq \sum_{i=0}^n \rho_{0,i}(x^i, y^i)$, where $\rho_{0,i} : \mathcal{X}_{0,i} \times \mathcal{Y}_{0,i} \rightarrow [0, \infty)$, is a sequence of $\mathcal{B}(\mathcal{X}_{0,i}) \times \mathcal{B}(\mathcal{Y}_{0,i})$ -measurable distortion functions, and let $\vec{Q}_{0,n}(D)$ (assuming is non-empty) denotes the average distortion or fidelity constraint defined by

$$\vec{Q}_{0,n}(D) \triangleq \left\{ \vec{q}_{0,i} \in \mathcal{M}_1(\mathcal{Y}_{0,i}), 0 \leq i \leq n : \frac{1}{n+1} \sum_{i=0}^n \int_{\mathcal{X}_{0,i}} \int_{\mathcal{Y}_{0,i}} \rho_{0,i}(x^i, y^i) \vec{q}_{0,i}(dy^i; x^i) \mu_{0,i}(dx^i) \leq D \right\}, \quad D \geq 0 \quad (19)$$

The causal rate distortion function associated with the causal product kernel $\vec{q}_{0,n} \in \vec{Q}_{0,n}(D)$ is defined by

$$\vec{R}_{0,n}(D) \triangleq \inf_{\vec{q}_{0,n} \in \vec{Q}_{0,n}(D)} \frac{1}{n+1} \mathbb{I}(\mu_{0,n}; \vec{q}_{0,n}) \quad (20)$$

while its operational meaning can be established via $\limsup_{n \rightarrow \infty} \vec{R}_{0,n}$.

Clearly, $\vec{R}_{0,n}(D)$ is characterized by minimizing directed information or equivalently $\mathbb{I}(\mu_{0,n}; \vec{q}_{0,n})$ over the causal product measure $\vec{q}_{0,n} \in \vec{Q}_{0,n}(D)$.

Lemma 2.6: $\vec{q}_{0,n} \in \mathcal{M}_1(\mathcal{Y}_{0,n})$ is uniquely determined by $\{q_i \in \mathcal{Q}_i(\mathcal{Y}_i; \mathcal{Y}_{0,i-1} \times \mathcal{X}_{0,i})\}_{i=0}^n$ and vice-versa, $P - a.s.$ Proof. For densities this result is derived in [15].

III. EXISTENCE OF OPTIMAL CAUSAL PRODUCT RECONSTRUCTION KERNEL

In this section, appropriate topologies and function spaces are employed to show existence of the minimizing causal product kernel in (20). In the process we also show existence for $R_{0,n}(D)$.

A. Abstract Spaces

Let $BC(\mathcal{Y}_{0,n})$ denote the vector space of bounded continuous real valued functions defined on the Polish space $\mathcal{Y}_{0,n}$. Furnished with the sup norm topology, this is a Banach space. The topological dual of $BC(\mathcal{Y}_{0,n})$ denoted by $(BC(\mathcal{Y}_{0,n}))^*$ is isometrically isomorphic to the Banach space of finitely additive regular bounded signed measures on $\mathcal{Y}_{0,n}$ [7], denoted by $M_{rba}(\mathcal{Y}_{0,n})$. Let $\Pi_{rba}(\mathcal{Y}_{0,n}) \subset M_{rba}(\mathcal{Y}_{0,n})$ denote the set of regular bounded finitely additive probability measures on $\mathcal{Y}_{0,n}$. Clearly if $\mathcal{Y}_{0,n}$ is compact, then $(BC(\mathcal{Y}_{0,n}))^*$ will be isometrically isomorphic to the space of countably additive signed measures, as in [3]. Denote by $L_1(\mu_{0,n}, BC(\mathcal{Y}_{0,n}))$ the space of all $\mu_{0,n}$ -integrable functions defined on $\mathcal{X}_{0,n}$ with values in $BC(\mathcal{Y}_{0,n})$, so that for each $\phi \in L_1(\mu_{0,n}, BC(\mathcal{Y}_{0,n}))$ its norm is defined by

$$\|\phi\|_{\mu_{0,n}} \triangleq \int_{\mathcal{X}_{0,n}} \|\phi(x^n)(\cdot)\|_{BC(\mathcal{Y}_{0,n})} \mu_{0,n}(dx^n) < \infty$$

The norm topology $\|\phi\|_{\mu_{0,n}}$, makes $L_1(\mu_{0,n}, BC(\mathcal{Y}_{0,n}))$ a Banach space, and it follows from the theory of “lifting” [10] that the dual of this space is $L_\infty^w(\mu_{0,n}, M_{rba}(\mathcal{Y}_{0,n}))$, denoting the space of all $M_{rba}(\mathcal{Y}_{0,n})$ valued functions $\{q\}$ which are weak*-measurable in the sense that for each $\phi \in BC(\mathcal{Y}_{0,n})$, $x^n \rightarrow q_{x^n}(\phi) \triangleq \int_{\mathcal{Y}_{0,n}} \phi(y^n) q(dy^n; x^n)$ is $\mu_{0,n}$ -measurable and $\mu_{0,n}$ -essentially bounded.

B. Weak*-Compactness and Existence

Define an admissible set of stochastic kernels associated with classical rate distortion function by

$$Q_{ad} \triangleq L_\infty^w(\mu_{0,n}, \Pi_{rba}(\mathcal{Y}_{0,n})) \subset L_\infty^w(\mu_{0,n}, M_{rba}(\mathcal{Y}_{0,n}))$$

Clearly, Q_{ad} is a unit sphere in $L_\infty^w(\mu_{0,n}, M_{rba}(\mathcal{Y}_{0,n}))$. For each $\phi \in L_1(\mu_{0,n}, BC(\mathcal{Y}_{0,n}))$ we can define a linear functional on $L_\infty^w(\mu_{0,n}, M_{rba}(\mathcal{Y}_{0,n}))$ by

$$\ell_\phi(q_{0,n}) \triangleq \frac{1}{n+1} \int_{\mathcal{X}_{0,n}} \left(\int_{\mathcal{Y}_{0,n}} \phi(x^n, y^n) q_{0,n}(dy^n; x^n) \right) \mu_{0,n}(dx^n)$$

This is a bounded, linear and weak*-continuous functional on $L_\infty^w(\mu_{0,n}, M_{rba}(\mathcal{Y}_{0,n}))$. For $d_{0,n} : \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n} \rightarrow [0, \infty)$ measurable and $d_{0,n} \in L_1(\mu_{0,n}, BC(\mathcal{Y}_{0,n}))$ the distortion constraint set of the classical rate distortion function is

$$Q_{0,n}(D) \triangleq \{q \in Q_{ad} : \frac{1}{n+1} \ell_{d_{0,n}}(q_{0,n}) \leq D\}$$

It can be shown that $Q_{0,n}(D)$ is bounded and weak*-closed subset of Q_{ad} and hence weak*-compact (Compactness of Q_{ad} follows from Alaoglu’s Theorem [7],[12]).

Next, we define the set of causal product kernels as follows.

$$\vec{\Pi}_{rba}(\mathcal{Y}_{0,n}) = \left\{ \vec{q}_{0,n}(dy^n; x^n) \triangleq \otimes_{i=1}^n q_i(dy_i; y^{i-1}, x^i) : q_i(dy_i; y^{i-1}, x^i) \in \Pi_{rba}(\mathcal{Y}_i), i \in \mathbb{N}^n \right\}$$

where $L_\infty^w(\mu_{0,n}, \vec{\Pi}_{rba}(\mathcal{Y}_{0,n}))$ denotes the space of all $\vec{\Pi}_{rba}(\mathcal{Y}_{0,n})$ valued functions $\{\vec{q}\}$ which are weak*-measurable in the sense that for each $\phi \in BC(\mathcal{Y}_{0,n})$, $x^n \rightarrow \vec{q}_{x^n}(\phi) \triangleq \int_{\mathcal{Y}_{0,n}} \phi(y^n) \vec{q}(dy^n; x^n)$ is $\mu_{0,n}$ -measurable and $\mu_{0,n}$ -essentially bounded.

Define the admissible set of causal product stochastic kernels associated with the causal rate distortion function by

$$\vec{Q}_{ad} \triangleq L_\infty^w(\mu_{0,n}, \vec{\Pi}_{rba}(\mathcal{Y}_{0,n}))$$

Clearly, $\vec{Q}_{ad} = \{q_{0,n} \in Q_{ad} : q_{0,n}(dy^n; x^n) = \vec{q}_{0,n}(dy^n; x^n)\}$. For $d_{0,n} : \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n} \rightarrow [0, \infty)$ which is measurable and $d_{0,n} \in L_1(\mu_{0,n}, BC(\mathcal{Y}_{0,n}))$ the distortion constraint of causal rate distortion function is

$$\begin{aligned} \vec{Q}_{0,n}(D) &\triangleq \left\{ \vec{q}_{0,n} \in \vec{Q}_{ad} : \right. \\ &\left. \frac{1}{n+1} \ell_{d_{0,n}}(\vec{q}_{0,n}) \triangleq \int_{\mathcal{X}_{0,n}} \left(\int_{\mathcal{Y}_{0,n}} d_{0,n}(x^n, y^n) \right. \right. \\ &\left. \left. \vec{q}_{0,n}(dy^n; x^n) \right) \mu_{0,n}(dx^n) \leq D \right\} \end{aligned}$$

Assumptions 3.1: We make the following assumptions.

- 1) The set \vec{Q}_{ad} is weak*-closed.
- 2) The set $\vec{Q}_{0,n}(D)$ is non-empty.

Lemma 3.2: Suppose Assumptions 3.1 hold. Let $\mathcal{X}_{0,n}, \mathcal{Y}_{0,n}$ be two Polish spaces and $d_{0,n} : \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n} \rightarrow [0, \infty]$, a measurable, non-negative, extended real valued function, such that $d_{0,n} \in L_1(\mu_{0,n}, BC(\mathcal{Y}_{0,n}))$. For any $D \in [0, \infty)$, the set $\vec{Q}_{0,n}(D)$ is weak*-compact.

Proof. By Assumptions 3.1, \vec{Q}_{ad} is a weak*-closed, hence as a subset of a weak*-compact set Q_{ad} it is weak*-compact. Also, under assumptions 3.1, $\vec{Q}_{0,n}(D)$ is bounded and weak*-closed and hence it is weak*-compact (as a weak*-closed subset of the weak*-compact set \vec{Q}_{ad}).

Theorem 3.3: Under Assumptions 3.1, $\vec{R}_{0,n}(D)$ has a minimum.

Proof. Follows from Lemma 3.2 and the lower semi-continuity of $\mathbb{I}(\mu_{0,n}; \cdot)$ on \vec{Q}_{ad} .

IV. NECESSARY CONDITIONS OF OPTIMALITY OF CAUSAL PRODUCT RATE DISTORTION FUNCTION

In this section the form of the optimal causal product reconstruction kernels is derived. The method is based on calculus of variations on the space of measures [9].

Theorem 4.1: Suppose $\mathbb{I}_{\mu_{0,n}}(\vec{q}_{0,n}) \triangleq \mathbb{I}(\mu_{0,n}; \vec{q}_{0,n})$ is well defined for every $\vec{q}_{0,n} \in L_\infty^w(\mu_{0,n}, \vec{\Pi}_{rba}(\mathcal{Y}_{0,n}))$ possibly taking values from the set $[0, \infty]$. Then $\vec{q}_{0,n} \rightarrow$

$\mathbb{I}_{\mu_{0,n}}(\vec{q}_{0,n})$ is Gateaux differentiable at every point in $L_{\infty}^w(\mu_{0,n}, \vec{\Pi}_{rba}(\mathcal{Y}_{0,n}))$, and the Gateaux derivative at the point $\vec{q}_{0,n}^0$ in the direction $\vec{q}_{0,n} - \vec{q}_{0,n}^0$ is given by

$$\begin{aligned} & \delta \mathbb{I}_{\mu_{0,n}}(\vec{q}_{0,n}^0; \vec{q}_{0,n} - \vec{q}_{0,n}^0) \\ &= \int_{\mathcal{X}_{0,n}} \int_{\mathcal{Y}_{0,n}} \log \left(\frac{\vec{q}_{0,n}^0(dy^n; x^n)}{\nu_{0,n}^0(dy^n)} \right) \\ & (\vec{q}_{0,n} - \vec{q}_{0,n}^0)(dy^n; x^n) \mu_{0,n}(dx^n) \end{aligned}$$

where $\nu_{0,n}^0 \in \mathcal{M}_1(\mathcal{Y}_{0,n})$ is the marginal measure corresponding to $\vec{q}_{0,n}^0 \otimes \mu_{0,n}(dx^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n} \times \mathcal{X}_{0,n})$.

Proof. The proof is based on the fact that the causal product stochastic kernel $\vec{q}_{0,n}$ is used to show the existence of Gateaux Differential [9] rather than for individual causal stochastic kernel $q_i(dy_i; y^{i-1}, x^i)$, $i \in \mathbb{N}^n$ •

The constrained problem defined by (20) can be reformulated using Lagrange multipliers as follows (equivalence of constrained and unconstrained problems follows from [9]).

$$\begin{aligned} \vec{R}_{0,n}(D) &= \inf_{\vec{q}_{0,n} \in \vec{\mathcal{Q}}_{ad}} \left\{ \frac{1}{n+1} \mathbb{I}(\mu_{0,n}; \vec{q}_{0,n}) \right. \\ & \left. - s(\ell_{d_{0,n}}(\vec{q}_{0,n}) - D) \right\} \end{aligned} \quad (21)$$

and $s \in (-\infty, 0]$ is the Lagrange multiplier.

Theorem 4.2: Suppose $d_{0,n}(x^n, y^n) = \sum_{i=0}^n \rho_{0,i}(x^i, y^i)$ and the assumptions of Lemma 3.2 hold. The infimum in (21) is attained at $\vec{q}_{0,n}^* \in L_{\infty}^w(\mu_{0,n}, \vec{\Pi}_{rba}(\mathcal{Y}_{0,n}))$ given by

$$\vec{q}_{0,n}^*(dy^n; x^n) = \otimes_{i=0}^n \frac{e^{s\rho_i(x^i, y^i)} \nu_i^*(dy^i; y^{i-1})}{\int_{\mathcal{Y}_i} e^{s\rho_i(x^i, y^i)} \nu_i^*(dy_i; y^{i-1})} \quad (22)$$

and $\nu_i^*(dy_i; y^{i-1}) \in \mathcal{Q}(\mathcal{Y}_i; \mathcal{Y}_{0,i-1})$. The causal rate distortion function is given by

$$\begin{aligned} \vec{R}_{0,n}(D) &= sD - \frac{1}{n+1} \sum_{i=0}^n \int_{\mathcal{X}_{0,i} \times \mathcal{Y}_{0,i-1}} \\ & \log \left(\int_{\mathcal{Y}_i} e^{s\rho_i(x^i, y^i)} \nu_i^*(dy_i; y^{i-1}) \right) \\ & \vec{q}_{0,i-1}^*(dy^{i-1}; x^{i-1}) \otimes \mu_{0,i}(dx^i) \end{aligned} \quad (23)$$

If $\vec{R}_{0,n}(D) > 0$ then $s < 0$ and

$$\frac{1}{n+1} \sum_{i=0}^n \int_{\mathcal{X}_{0,i}} \int_{\mathcal{Y}_{0,i}} \rho_{0,i}(x^i, y^i) \vec{q}_{0,i}^*(dy^i; x^i) \mu_{0,i}(dx^i) = D$$

Proof. The fully unconstrained problem of (21) is obtained by introducing another Lagrange multiplier. Using this and Theorem 4.1 we obtain (22) and (23) •

V. PROPERTIES OF CAUSAL RATE DISTORTION FUNCTION

In this section, we present some important properties of the causal rate distortion function as it is defined in (20).

Theorem 5.1:

- 1) $\vec{R}_{0,n}(D)$ is a convex, non-increasing function of D
- 2) If $\rho_{0,i} \in L^1(\pi_{0,i})$ then
 - a) $\vec{R}_{0,n}(\frac{1}{n+1} \sum_{i=0}^n E_{\pi_{0,i}}(\rho_{0,i})) = 0$;

- b) $\vec{R}_{0,n}(D)$ is non-increasing for $D \in [0, D_{max}]$ where $D_{max} = \frac{1}{n+1} \sum_{i=0}^n E_{\pi_{0,i}}(\rho_{0,i})$ and $\vec{R}_{0,n}(D) = 0$ for any $D \geq D_{max}$
- 3) $\vec{R}_{0,n}(D) > 0$ for all $D < D_{max}$ and $\vec{R}_{0,n}(D) = 0$ for all $D \geq D_{max}$, where

$$D_{max} = \min_{\{y^n\} \in \mathcal{Y}_{0,n}} \frac{1}{n+1} \sum_{i=0}^n \int_{\mathcal{X}_{0,i}} \rho_{0,i}(x^i, y^i) \mu_{0,i}(dx^i)$$

if such a minimum exists.

Proof. Omitted due to space limitation.

VI. CONCLUSION AND FUTURE WORK

A. Conclusion

The solution of the causal rate distortion function subject to a reproduction kernel which is a product of causal kernels is presented, on abstract alphabets. Some of its properties are also presented. It is believed that the optimal reconstruction kernel as a product of causal kernels has several implications in applications where causality of the decoder as a function of the source is of concern.

B. Future Work

Examples are currently under investigation, and will be presented at the final version of the paper.

VII. APPENDIX

REFERENCES

- [1] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice Hall, Englewood Cliffs, NJ, 1971.
- [2] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley & Sons, 1991.
- [3] I. Cs        , "On an extremum problem of information theory", *Studia Scientiarum Mathematicarum Hungarica*, vol. 9, pp. 57–71, 1974.
- [4] D. L. Neuhoff and R. Kent Gilbert, "Causal Source Codes", *IEEE Transactions on Information Theory*, vol. IT-28, No.5, pp. 701–713, 1982.
- [5] S. Tatikonda, "Control Over Communication Constraints", *PhD Dissertation*, M.I.T., Cambridge, MA, 2000.
- [6] J. Massey, "Causality, Feedback and Directed Information", in the *IEEE International Symposium on Information Theory and its Applications*, pp. 303–305, Nov. 27–30, Hawaii, U.S.A.1990.
- [7] N. Dunford and J. T. Schwartz, *Linear Operators, Part I: General Theory*. Interscience Publishers, Inc., New York, 1958.
- [8] R. M. Gray, *Entropy and Information Theory*. Springer-Verlag, 1990.
- [9] D. G. Luenberger, *Optimization by Vector Space Methods*. John Wiley & Sons, 1969.
- [10] A. Ionescu Tulcea & C. Ionescu Tulcea, *Topics in the Theory of Lifting*, Springer Verlag, Berlin, Heidelberg, New York, 1969.
- [11] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the theory of Large Deviations*. John Wiley & Sons, 1997.
- [12] W. Rudin, *Functional analysis*. McGraw-Hill, 1991.
- [13] H.B.Maynard, *A Radon-Nikodym Theorem for Finitely Additive Bounded Measures*, *Pacific Journal of Mathematics*, 83(2), 1979, pp. 401-413.
- [14] F. Rezaei, N. U. Ahmed and C. D. Charalambous, *Rate Distortion Theory for General Sources With Potential Application to Image Compression*, *International Journal of Applied Mathematical Sciences*, vol. 3 No. 2, 2006, pp. 141-165.
- [15] H. H. Permuter, T. Weissman, A. Goldsmith, "Finite State Channels with Time-Invariant Deterministic Feedback", *IEEE Transactions on Information Theory*, vol.IT-55, No. 2, pp. 644-662, February 2009.